Stat 414 – Day 31
Longitudinal data (Ch. 15)

Previously: Have data nested within groups. Want to include the grouping variable in the model. Including it as random intercepts gives us a multilevel model, which has advantages including

- Allows separation of within group and between group variation
- Allows for inclusion of Level 1 and Level 2 variables
- Induces/Estimates within group correlation
- Including random slopes models heterogenous responses (Level 2)
- Including cross-level interactions can explain variation in slopes (Level 2 equation)
- Does not require equal group sizes/handles missing values well

Multilevel models are especially helpful for "longitudinal data" (e.g., repeat observations on the same individual over time). Typically with longitudinal data, we want to focus on changes over time and the effect of Level 2 covariates. (We've actually already been looking at repeated measures data (!) but you will often see some different terminology come up.)

Example: Data were collected by the Minnesota Department of Education for all Minnesota schools during the years 2008-2010 to compare charter and non-charter schools. School performance is measured by the mean score on the math portion of the Minnesota Comprehensive Assessment (MCA-II) data for the 6th grade students enrolled in 618 different Minnesota schools during the years 2008, 2009, and 2010. (MCA test scores for sixth graders are scaled to fall between 600 and 700, where scores above 650 for individual students indicate "meeting standards." Thus, schools with averages below 650 will often have increased incentive to improve their scores the following year.)

(a) Identify the level 1 and level 2 units. (What are i and j?)

Initial data exploration

(b) First, we want to explore how MCA math test scores relate to important Level 2 variables (e.g., percentage of students with free and reduced lunch). This can be done using the data values for all three years or by averaging the data values for the three years into one number or only using 2010 values. Give a brief pro/con of these approaches.

For the second approach, open the "wide format" of the data (chart.wide.txt, this includes three columns for the three time points for each school) and use the SchoolAvg variable as the response. Examine the associations of these variable with several of the Level 2 variables.

(c) Which variables seems most useful in predicting the average math score? Does the correlation matrix make sense in this context?

Now open the "long format" of the data (chart.long.txt). Create two visual representations of math scores vs. time for the first 20 schools:

- separate graphs for each school
- connecting lines or smoothers for each school overlaid on same graph (i.e., "spaghetti plot")

(d) Explain what *year08* represents.

(e) Does it look like we will want to include random intercepts? (Meaning?) Does it look like we will want to include random slopes? (Meaning?)

(f) Produce a graph of the Math scores vs. year, separated by the charter (charter = 1) vs. public (charter = 0) schools. What do you learn?

Modeling

Fit the *unconditional growth model* (time is only Level 1 variable, we haven't "conditioned" or "controlled" for any other possible covariates): multilevel model with year08, random intercepts and slopes. (Be sure to use schoolnum, which are unique, not school name):

Unconditional growth model:

$$y_{ij} = \beta_{00} + \beta_{10} time_{ij} + u_{0j} + u_{1j} time_{ij} + \epsilon_{ij}$$

$$where\ \epsilon_{ij} \sim N(0, \sigma^2 \boldsymbol{I})$$

```
Random effects:
 Groups    Name        Variance Std.Dev. Corr
 schoolnum (Intercept) 39.4410  6.2802
           year08       0.1105  0.3325   0.72
 Residual               8.8200  2.9699
Number of obs: 1733, groups:  schoolnum, 618

Fixed effects:
             Estimate Std. Error t value
(Intercept) 651.40766    0.27934 2331.96
year08        1.26495    0.08997   14.06
```

(g) Describe what this model is doing. Interpret the variance components. How would you determine the percentage of within-school variation that is explained by the linear increase over time? What assumption does this model make about the "occasion-specific" residuals? Does that seem like a reasonable assumption in this context?